Minihomework 2

Roberto Navigli and Alessandro Scirè

Extractive Summarization

- We just created a dataset made up of **100 short stories and fairy tales** + their **summaries**.
    - In English
- We will assign **2 short stories and 2 fairy tales** for a total of less than 1,000 sentences
- The task will be to select a **small number of sentences** that summarize each text
    - Carried out by following the summary provided
- You will use and evaluate a **state-of-the-art extractive summarization system**
    - E.g. from Hugging Face
- You will provide a **short analysis/report** of the small experiment
- **Research track 2**: the best works will be selected to participate in a short ACL paper where we will study the extent to which a text can be summarized extractively in a more or less objective manner

Extractive Summarization - Guidelines

1. Read the summary text
2. Open the story and summary texts and keep them in the same screen
3. Read the story text and highlight the **salient** sentences
    a. A sentence is **salient** if it **contains part of the summary information**
4. Once you finish, copy the highlighted sentences in a separate .txt (plain text) file
5. Read the outcome of your selection and apply modifications if needed

If you encounter any **critical issue concerning the text quality**, please report to ALL of us by email!
{scire, navigli}@diag.uniroma1.it

Extractive Summarization - Guidelines

- Keep in mind the following criteria to judge if your selection is acceptable:
  - **Relevance:** The selected sentences are **important** for the plot
  - **Consistency:** The selected sentences don't introduce any factual inconsistency. This can happen, for example, with ambiguous pronouns.
    - Ex: 1. The horse is running.
      2. The cat is hiding behind the rock.
      3. It is a very nice feline.
  - **Coherence:** The selected sentences (in yellow) follow a fluent logical order so that the produced text can be easily read/understood by a human
  - **As few sentences as possible to make the story coherent and fluent**
- To sanity check the outcome of your annotation, you can provide the result to another person (NOT enrolled in this course!) and see if (s)he is able to grasp the salient content and rationale of the story only by reading your sentences

Extractive Summarization - Guidelines

- You cannot select parts of sentences. **ONLY entire sentences** (denoted by a unique id) **are allowed**
  - Note that sentences have been split automatically, so please select multiple contiguous rows if a sentence was erroneously broken into multiple rows
- Feel free to add sentences to resolve pronoun ambiguities:

  Ex: 1. The horse is running.
  2. The cat is hiding behind the rock
  3. It is a very nice feline

  Given that sentence 3 is annotated, sentence 2 must be highlighted too, in order to resolve the ambiguity of the pronoun ‘it’.

  - only in cases like this, you are free to add sentences that are not **salient** (see slide 3)

Extractive Summarization - Code

- After the annotation task, select a state-of-the-art extractive summarization system.
    - E.g., from Huggingface
- Execute the system **on the whole dataset (the whole collection, <u>not only your stories</u>)** and obtain the set of predicted sentences

Extractive Summarization –Evaluation & Report

- Report the **issues/challenges** encountered during the annotation

- **HUMAN-SYSTEM comparison:**
  Evaluate the extractive summarization system predictions by comparing them with your produced annotation (if the number of sentences can be provided, do provide your own number of sentences; else let it free).
  - You can use standard summarization evaluation metrics, i.e. **ROUGE-L**, as well as **F1**, **precision** and **recall** between the manually selected sentence ids and the predicted ones.
  - Some qualitative comparison

- **HUMAN-HUMAN comparison:**
  Compare your manual annotations with the ones made by another student on the same text.

Write a **short report** containing the results of the comparisons, providing key insights. Qualitative examples will be appreciated :)

What you will receive

- A directory containing texts and stories summaries; you can find them here!
- You will receive an email containing the filepaths of the stories to annotate.
- PLEASE, **do NOT share your assigned stories with other students**
    - This is to avoid any cooperation between you, which will affect the validity of the experimental results

What we expect from you

- Submit a **.zip** file containing:
  - **results.tsv** with the predictions of the extractive summarization system.
    - We expect a **.tsv** populated with the following columns:
      - **DOCUMENT_ID:** (string), corresponding to the filename of the story without extension, e.g. bn:01899260n_The Raven
      - **SENTENCES_IDS:** comma separated list of ids, corresponding to the identifiers of the predicted sentences. E.g. 1,2,5,7,8
  - A **.txt** for each assigned story, containing the **selected sentences ONLY**
    - Name the files using the following convention: {DOCUMENT_ID}.txt, e.g., bn:01899260n_The Raven.txt
    - **On Windows, you are allowed to replace : in filenames with space or _**
  - **extractive_summarization.ipynb**: a Colab notebook with the code to execute the extractive summarization system on the whole dataset.
  - **report.pdf**: a short report with the issues/challenges encountered during the annotation and key insights about the HUMAN-HUMAN and HUMAN-SYSTEM comparisons (see slide 7)
    - Use the ACL Overleaf template to generate the pdf

What we expect from you

- Name the **.zip** file **`lastname_studentid_minihw2.zip`**:
  - Ex: Luigi D’Andrea will submit a file named **`dandrea_1234567_minihw2.zip`**
  - If you are unsure which name to put, use the one in your institutional email account

⚠ Any violation to the above convention will invalidate your submission!

What you can and cannot do

- You can use whatever library (imported from your notebook)
- You **cannot** plagiarize!!!
    - If you do it, all students involved will be excluded

Submission

- You have to submit the .zip file through the submission form on Google Classroom. You will be asked to fill a form with the requested information and attach the .zip file.

- **Deadline (all files but the report): December 8th (AoE)**

- You will receive the other summaries of your stories made by a colleague of yours on **December 9th**
    - However, we recommend that you finalize the report on the issues/challenges encountered during the annotation task

- **Deadline (report): December 15th (AoE)**

    - Link to be provided!!!

That’s not all

If your work is novel, interesting and original, we will gladly invite you to work together with us to extended on a fully-fledged paper for **TOP-TIER INTERNATIONAL CONFERENCE**!

Just over the last 12 months, the Sapienza NLP group published more than a dozen of papers!

Questions/Issue?

If you have a question that may interest your colleagues, **please ask it on Google Classroom.**

Otherwise, for personal or other questions, email **ALL** of us (but please, only reach for things that can't be asked on the Google Classroom).

If you encounter any **critical issue concerning the text quality**, please report to us by email !

Our emails are:

{scire, orlando, bonomo}@diag.uniroma1.it

Good Luck!!!

Sapienza NLP